

John Heist, Fotolia

Las herramientas del proyecto Simile para la web semántica

# LAS HERRAMIENTAS DEL MAÑANA

El lanzamiento del proyecto Simile inicia la web semántica con una colección de herramientas que extienden la información semántica a los sitios web existentes. **POR OLIVER FROMMEL**

Una simple búsqueda en Google nos enseña lo tonto que es la web. Si buscamos la solución a un problema con Linux, muy probablemente encontremos a un montón de usuarios con el mismo, pero no la solución en sí. El error es que Google simplemente evalúa la coincidencia con cualesquiera palabras que el usuario ingresa en el momento de describir el problema.

Un buscador de internet tradicional no analiza la estructura del documento o de la conversación. Esto hace que, por ejemplo, buscando la palabra 'Linux', encontremos artículos que poco tienen que ver con Linux.

Google sólo realiza un análisis rudimentario del lenguaje, corrige erratas o frases mal escritas basándose en métodos estadísticos conocidos como

*Ngrams*. De todas formas, es incapaz de interpretar el significado de las palabras clave de la búsqueda. Por ejemplo, no distingue entre banco como institución financiera o banco como sitio donde sentarse.

Hay gente que se pregunta si las máquinas serán capaces o no de ordenar basándose realmente en la terminología de la lengua humana; de cualquier modo, la comunidad de desarrolladores web de Simile está empeñada en que esto deje de ser una duda y pase a ser una aseveración.

El proyecto Simile desarrolla herramientas para la web semántica. Los visionarios de esta web semántica [1] esperan que las herramientas como las que proporciona Simile hagan más inteligente, y por tanto más útil, el procesamiento por parte de las máquinas.

## Problemas

Los desarrolladores web cuentan con un conjunto de sofisticadas técnicas para servir un contenido web dinámico tratado como datos por las aplicaciones cliente. Este paradigma es, por un lado, costoso en términos de esfuerzo, y por otro, demasiado limitado para las necesidades de la web semántica. Uno de los problemas se debe al hecho de que este tipo de funcionalidad debe ser programada cuidadosamente en el momento de la elaboración del propio sitio web, lo que podría estar bien si se está contruyendo un sitio web nuevo, pero no si se quieren obtener los beneficios de la automatización con sitios web estáticos ya existentes. El otro problema es que una aplicación de servidor convencional ha de escribirse con un conocimiento muy detallado sobre

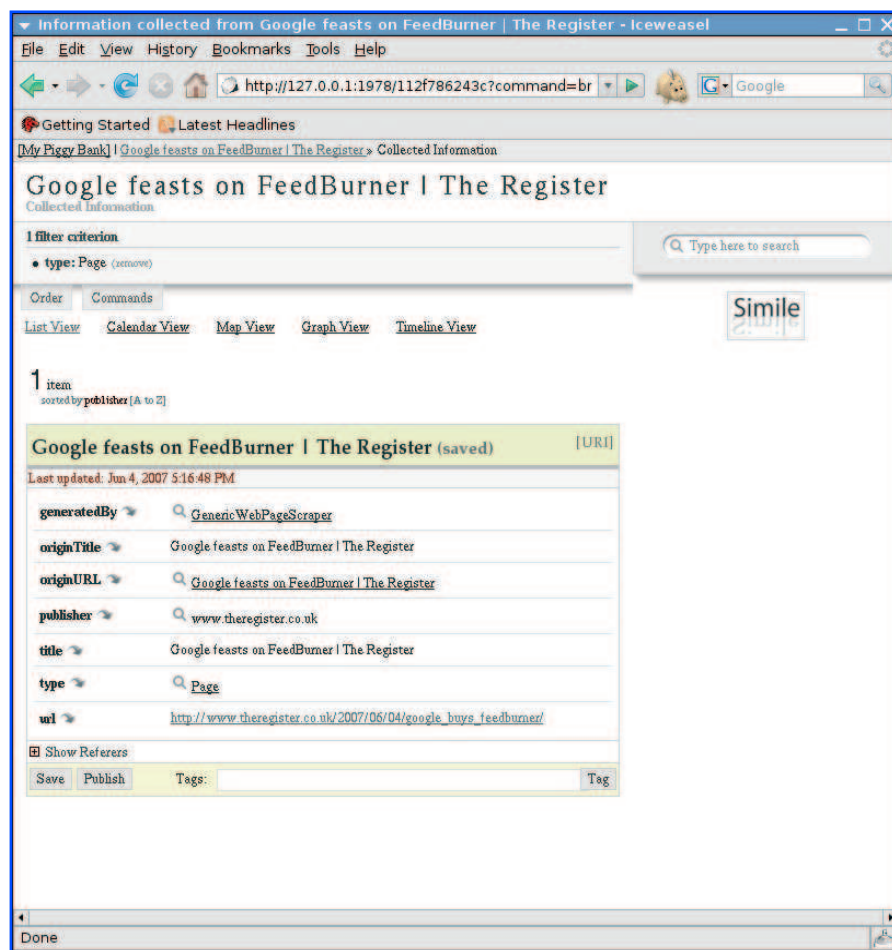


Figura 1: Sin un extractor, Piggy Bank sólo descubre algunos retazos de información en las páginas HTML.

qué hará el cliente con los datos que va a recibir

La web semántica resuelve ambos conflictos, adoptando la premisa de que si la información web tiene un significado muy parecido al que tendría en el lenguaje real, ésta se vuelve más interpretable y automatizable, sin que sea necesaria una coordinación exagerada entre el servidor y el cliente.

Para que la web semántica siga avanzando, las nuevas tecnologías deben incluir la metainformación necesaria y proveer unos medios simples para ligar esta información semántica a los sitios web existentes. El proyecto Simile, patrocinado por el MIT (Massachusetts Institute of Technology), desarrolla herramientas software que podrían algún día ayudar a suavizar la transición hacia este tipo de web. Simile significa *Semantic Interoperability of Metadata In unLike Environments*. Las incluidas en el juego de herramientas de Simile (Tabla 1) han sido diseñadas para mostrar, extraer,

asociar y manipular información semántica.

Hasta ahora, las técnicas relacionadas con la web semántica no han subsistido, en parte debido a la lentitud del proceso de estandarización del consorcio W3, pero también debido a la ingente cantidad de tecnologías y estándares diferentes usados en la práctica. Probablemente, el formato más popular de almacenamiento de datos semánticos sea RDF (Resource Description Framework). RDF se usa ya en algunos RSS populares de noticias [2], siendo la base también de muchas de las herramientas del proyecto Simile.

## Formato RDF

El formato RDF ha sido diseñado para estructurar los datos de la web de modo que sean independientes de su formato. Su propósito es organizar estos datos para que puedan ser interpretados según su significado en vez de ser vistos como simples letras y palabras. RDF estructura los recursos con

expresiones compuestas por tres partes, conocidas como *triples*. Un triple refleja la forma clásica de una frase, que consiste también en tres partes:

- asunto
- predicado
- objeto

La introducción que Wikipedia hace al RDF [3] nos pone de ejemplo la frase *El cielo es azul*. Ésta puede expresarse como un triple de RDF con *El cielo* como asunto, *azul* como objeto y *es* como el predicado que relaciona a asunto y objeto.

En cualquier idioma, las posibilidades de estructuración de oraciones en formato RDF son infinitas. El propósito de la tecnología RDF no es elaborar un solo programa que navegue a través de todo el lenguaje natural, sino proporcionar medios sencillos, al propietario (o usuario) de una página web, de dar pistas sobre el tipo de contenido de la página. La intervención humana sigue siendo necesaria, pero con RDF se reduce a una forma simple y concisa, minimizando el caos con las páginas web existentes.

Las herramientas del proyecto Simile funcionan con la información RDF derivada de

- una definición explícita de las relaciones RDF creada por el dueño de la página web y referenciada en forma de enlace en la cabecera del HTML
- una herramienta especialmente diseñada para convertir el contenido de una página web al formato RDF

Para el propietario de un sitio web, expresar la información de la web por medio de una definición explícita hace que dicha información pueda ser integrada en herramientas de automatización personalizadas. Por ejemplo, el dueño de una cadena de hoteles, cuyo sitio web contiene un listado con los nombres y direcciones de sus hoteles, puede hacer que el webmaster de sus páginas cree un archivo RDF que contenga las mencionadas direcciones, de tal forma que las agencias de viajes puedan elaborar fácilmente herramientas que las sitúen en un mapa de Google.

RDF nos permite este tipo de automatización sin necesidad de reestructurar completamente el sitio web. La página web puede seguir mostrando el listado de direcciones y nombres de hoteles. El

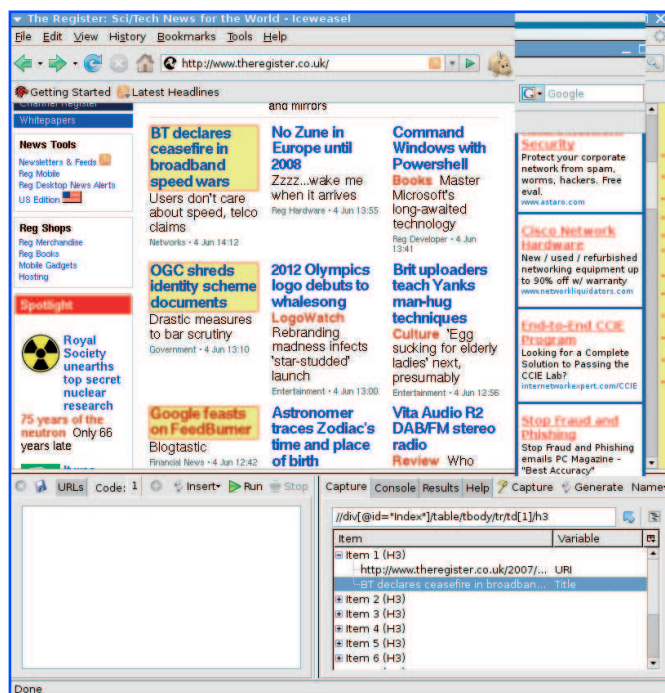


Figura 2: La extensión de Firefox, Solvent, ayuda a los usuarios a escribir sus propios extractores..

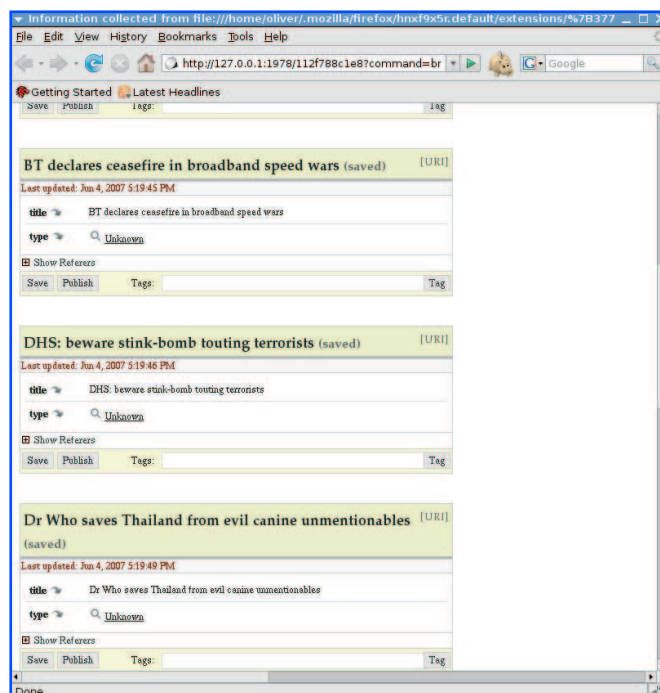


Figura 3: Piggy Bank es capaz de mostrar la información extraída por Solvent de un sitio web complementada con marcas semánticas..

único cambio necesario es la inclusión de una referencia a la definición RDF en la cabecera del fichero HTML.

La opción del analizador de contenidos requiere una menor intervención por parte del propietario de la página. De hecho, ni siquiera tiene por qué enterarse de que el visitante está automatizando la información. Podemos escribir nuestras aplicaciones de modo que procesen el texto de una página web y la dote de una estructura RDF. También puede ser el mismo propietario quien defina qué analizador puede usarse para sus páginas. El Listado 1

muestra el código de ejemplo proporcionado por el proyecto Simile para referenciar un RDF o un analizador en una cabecera de HTML.

Piggy Bank, en el núcleo del juego de herramientas de Simile, es una extensión para Firefox. Nos permite ver, organizar y combinar información RDF de diferentes fuentes, y lo podemos usar como gestor para distintos analizadores de contenidos. Otra herramienta del juego es Solvent, que nos ayuda a escribir nuestro propio analizador para la extracción de datos RDF de los sitios web.

Las herramientas de la web semántica, como Piggy Bank o Solvent, están aún en una fase experimental muy inicial. Este artículo nos deja vislumbrar algunas de las herramientas de Simile para esta forma de entender la web. Más información sobre el juego de herramientas de Simile en el sitio web del proyecto [4].

### Explorando Piggy Bank

Una vez en la página principal de Piggy Bank [5], sólo hay que pinchar en el enlace a una extensión de Firefox para instalarla. El instalador da por hecho que tenemos el plugin de Java.

En nuestro laboratorio, Piggy Bank no funcionaba con los plugins de JDK 1.4.2, o JDK 1.6.0, sólo con JDK 1.5.0. Para activar el plugin instalado con JDK debemos crear un enlace simbólico:

```
cd $HOME/.mozilla/plugins
ln -s /usr/java/jdk1.5.0_11/jre2
/plugin/i386/ns7
/libjavaplugin_oji.so
```

Si esto no funciona, un usuario de Firefox no puede esperar mucha más ayuda. Simplemente no se ejecutará el plugin de Simile en el navegador, aunque siempre podemos intentar ejecutar el navegador en modo desarrollo:

Herramienta	Descripción
Piggy Bank	Extensión de Firefox que almacena localmente los resultados de las peticiones Xpath
Solvent	Extensión auxiliar de Piggy Bank para extracción de datos
Semantic Bank	Servidor que almacena y centraliza la información de Piggy Bank de muchos usuarios
Welkin	Visualización de Gráficos RDF
Longwell	Navegador para datos RDF
Gadget	Inspector para juegos de datos XML de gran tamaño
Referee	Extractor de metadatos
Exhibit	Generador automático de páginas HTML a partir de una base de datos
Babel	Traductor de formatos; por ejemplo, de JSON a Exhibit
Fresnel	Vocabulario para el renderizado de RDF
HTTPTracer	Herramienta de monitorización de tráfico HTTP
Timeline	Widget DHTML para ordenar datos cronológicamente
RDFizer	Colección de herramientas para conversión RDF

### Listado 1: Mostrando Extractores y archivos RDF

```

01 <html>
02   <head>
03     ...
04     <link rel="alternate"
05       type="application/n3"
06       title="Screen scrapers'
07         information"
08       href="http://people.csail.mit.
09         edu/dfhuynh/foaf.rdf">
10     </head>
11     ...
12 </html>

```

firefox -P development

Si todo ha salido bien, debe aparecer un nuevo botón en el borde inferior de la ventana o bien el icono de un cerdito junto a la barra de direcciones, que iniciará el navegador de Piggy Bank.

La primera vez que ejecutamos el programa estará vacía la base de datos local. El nuevo elemento del menú *Herramientas* | *Piggy Bank* | *Collect and Browse* nos permite llenarla.

La extensión Piggy Bank busca recursos RDF en la página actual y los guarda. Si la página no tiene nada de información semántica, Piggy Bank ejecuta los analizadores de contenidos disponibles para tratar de recoger los

datos de la pantalla. Cada analizador debe escribirse especialmente para cada página, ya que depende en gran medida de la estructura del documento.

### Recolectando Información

La extensión Piggy Bank incluye tres analizadores listos para su uso, además de otros que hay en el sitio web. Dicho esto, se entiende que han de ser los usuarios quienes escriban sus propios analizadores, de lo contrario, Piggy Bank sólo muestra trozos de información que ha ido detectando automáticamente en páginas web estándar, como URLs o títulos (Figura 1).

Como apoyo a la escritura de nuestros propios analizadores, Simile

incluye otra extensión de Firefox que simplifica el proceso: Solvent. Tras completar su instalación, podemos ejecutarla pulsando sobre el icono con aspecto de aerosol que aparece en la parte inferior derecha de la ventana. Al hacerlo, Firefox abre dos ventanas más en el área que ocupa la ventana actual (Figura 2).

Para extraer la información de un sitio web pulsamos el botón *Capture* y el cursor adopta la forma de una mano. Podemos entonces pinchar sobre los elementos HTML que nos interesen; Solvent marcará con amarillo aquellos elementos sobre los que se encuentre el cursor. Al pinchar sobre ellos, la extensión selecciona automáticamente el resto de elementos que coinciden con la expresión *Xpath*, que podemos ver en la primera línea de la parte superior derecha de la ventana. Al mismo tiempo, Solvent muestra en la mitad inferior los elementos seleccionados. La flecha azul, a la derecha de la expresión *Xpath*, selecciona los elementos de la capa inmediatamente superior.

Tras identificar los elementos deseados de la página, el usuario debe añadir la información semántica, que obviamente no está en el código HTML. Para hacerlo, primero expandimos uno de los elementos seleccionados y luego pulsamos el botón *Name*. Este botón nos lleva a una lista de marcas semánticas predeterminadas, aunque también podemos definir nuestras propias marcas. En este caso deberíamos seleccionar los elementos: *URI* para la URL y *Title* para el título.

Es posible asignar etiquetas a través del menú; sus nombres pueden incluir una URL de modo que cumplan los estándares existentes. Simile saca la mayor parte de las etiquetas predefinidas del Dublin Core [6], el cual implementa una taxonomía usada, por ejemplo, por el formato EDF de OpenOffice para sus metadatos. El Dublin Core tiene un prefijo URL

```
http://purl.org/dc/elements/1.1/
```

Lo normal es que no obtengamos los resultados esperados usando estos extractores si no los asistimos manualmente. En última instancia,

### Microformatos

El confinamiento del lenguaje de marcado de las páginas HTML limita actualmente la web a una mera representación de una capa de texto. Bien sea un listado de personas o un listado de coches, las etiquetas de marcado de HTML siempre son iguales: *li*, *div*, o *td*.

Recientemente, en el contexto de la web 2.0, se han promocionado de forma exagerada los microformatos. Éstos intentan añadir algo de semántica a la web HTML incluyendo meta-información a los atributos de las clases de los elementos HTML. Por ejemplo:

```
<div class="Calle">
Calle del suspiro</div>
```

Como pueden imaginar, esto funciona mejor con un marcado minimalista que con páginas sobrecargadas de etiquetas. Por eso las aplicaciones principales de los microformatos son hoy por hoy entradas

de calendarios o tarjetas de negocios. Los usuarios pueden adjuntar distintas palabras clave (a las que también nos referimos como etiquetas) para sus datos en línea, como por ejemplo fotografías en flickr.com. El almacenamiento basado en servidores y las técnicas Ajax permiten la adición de etiquetas y el uso de las ya existentes.

Algunos observadores se refieren a los microformatos y la ontología como la base activa de la web semántica, ya que proporcionan un enfoque adecuado para la resolución de los problemas de esta web. Incluso algunos puristas, a pesar de no insistir en el uso de ontologías construidas por comités de estandarización, ven el potencial de los microformatos. Se puede considerar a la web 2.0, por tanto, como una especie de fase transicional en la que se integrarán las tecnologías semánticas cada vez más.

necesitaremos modificar los nombres de los elementos en el código. Para extraer la información de páginas con estructuras complejas necesitamos algo más que expresiones *Xpath*. En este caso, necesitaríamos procesar el

HTML nosotros mismos con algo de JavaScript.

### Generación de Códigos Extractores

Después de montar los elementos de la página y los metadatos podemos pulsar el botón *Generate* para crear nuestro código, que Solvent mostrará en la ventana, a nuestra izquierda. Otro clic sobre *Run* ejecuta el extractor sobre la página actual y muestra el resultado en formato RDF a nuestra derecha. El botón *Show Results* de Piggy Bank presenta los resultados en su navegador (Figura 3), pudiendo añadir ya las etiquetas a la información y guardar los datos localmente.

Como alternativa a un repositorio local, los usuarios pueden almacenar la información semántica en el servidor *Semantic Bank*, que forma parte también del proyecto Simile. El banco semántico ofrece a los usuarios la oportunidad de colaborar. Otras herramientas de Simile soportan la extracción de páginas mediante la línea de comandos,

convirtiendo el resultado de los datos RDF a otros formatos y mostrando dichos datos en orden cronológico o geográfico.

### Es Necesario Algo de Código

Ni siquiera las herramientas del proyecto Simile son capaces de automatizar completamente la elaboración de la web semántica a partir de las ofrendas heredadas de la web. Las estructuras existentes en los sitios web son demasiado diversas como para que esto ocurra. Incluso para los extractores de pantallas se necesitan conocimientos sobre *Xpath* y *JavaScript*.

Sólo con señalar y pinchar sobre un sitio web complejo no vamos a conseguir demasiado. Gracias a la variedad de herramientas, los usuarios ya pueden experimentar con las tecnologías semánticas sin tener que manipular el código RDF. Cabe esperar que de aquí a poco vayan apareciendo nuevos extractores y otras herramientas, al menos para sitios extremadamente populares. ■

RECURSOS
[1] La web semántica: "Una nueva forma de contenido Web significativa para las máquinas liberará una revolución de nuevas posibilidades", Tim Berners-Lee, James Hendler, y Ora Lassila, Scientific American.com, Mayo de 2001: <a href="http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&amp;catID=2">http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&amp;catID=2</a>
[2] Resource Description Format: <a href="http://www.w3.org/RDF">http://www.w3.org/RDF</a>
[3] RDF en la Wikipedia: <a href="http://es.wikipedia.org/wiki/RDF">http://es.wikipedia.org/wiki/RDF</a>
[4] Simile: <a href="http://simile.mit.edu">http://simile.mit.edu</a>
[5] Piggy Bank: <a href="http://simile.mit.edu/wiki/Piggy_Bank">http://simile.mit.edu/wiki/Piggy_Bank</a>
[6] Dublin Core: <a href="http://dublincore.org/">http://dublincore.org/</a>

SEGUNDO SEGUNDO SEGUNDO SEGUNDO SEGUNDO

# CONCURSO UNIVERSITARIO DE SOFTWARE LIBRE

SIGUE LA EVOLUCION DE LOS PROYECTOS DESDE:  
[WWW.CONCURSOSOFTWARELIBRE.ORG/PLANET](http://WWW.CONCURSOSOFTWARELIBRE.ORG/PLANET)

PLANET

Patrocinador Oro:

Patrocina:

Colaborador Principal:

Medios oficiales:

Universidades Colaboradoras:

Organismos Colaboradores:

Organiza: